

# Software Defined Networking @ESnet

Inder Monga

Chin Guok

OASCR SDN Workshop May 12<sup>th</sup> 2014



# Goals



- Highlight ESnet efforts in SDN
- Discuss future directions for R&E community

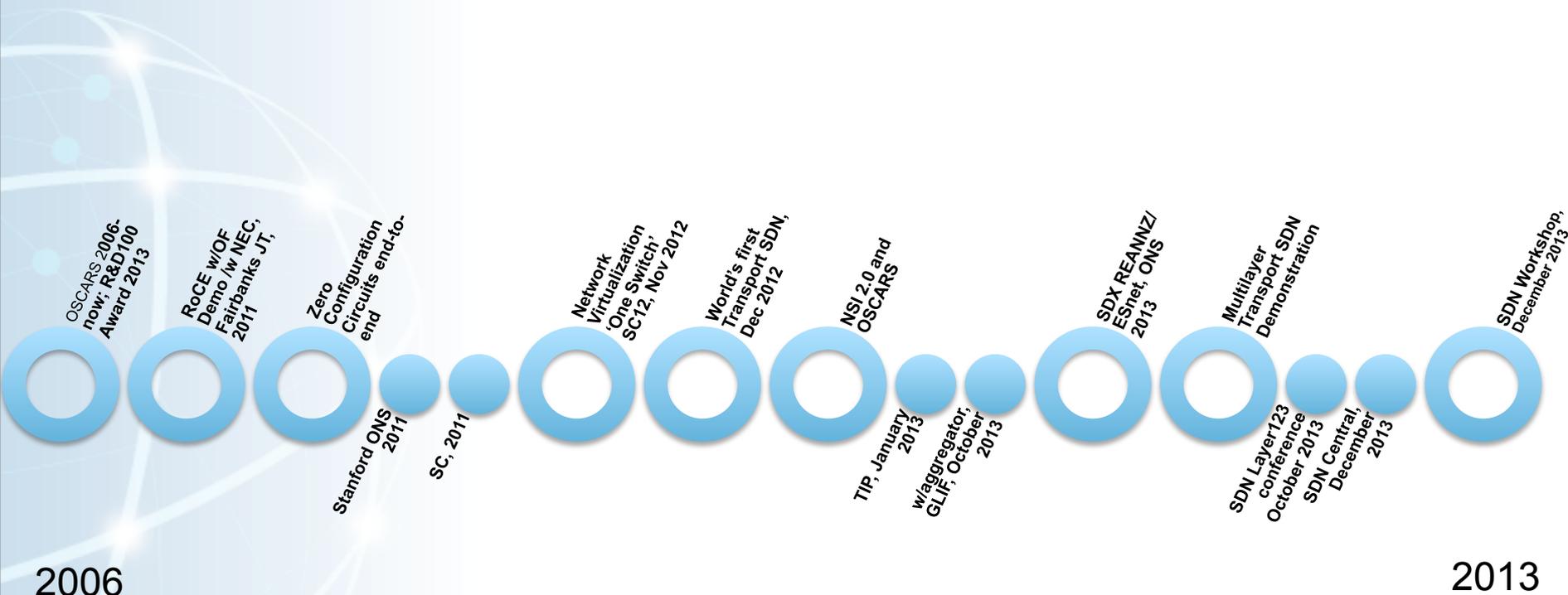
# Funding background



OASCR Research Funding: 1 year funding to investigate SDN for ESnet

Program Funding: Continue the research and prototyping work for SDN

# SDN Investigations @ ESnet brief timeline



ONS = Open Networking Summit  
TIP = Joint Techs with Univ. Hawaii  
RoCE= Remote DMA over Converged Ethernet  
GLIF = Global Lambda Integrated Facility  
SDN = Software Defined Networking

# Journey with Programmability

Joint Techs Summer 2011, Fairbanks, Alaska

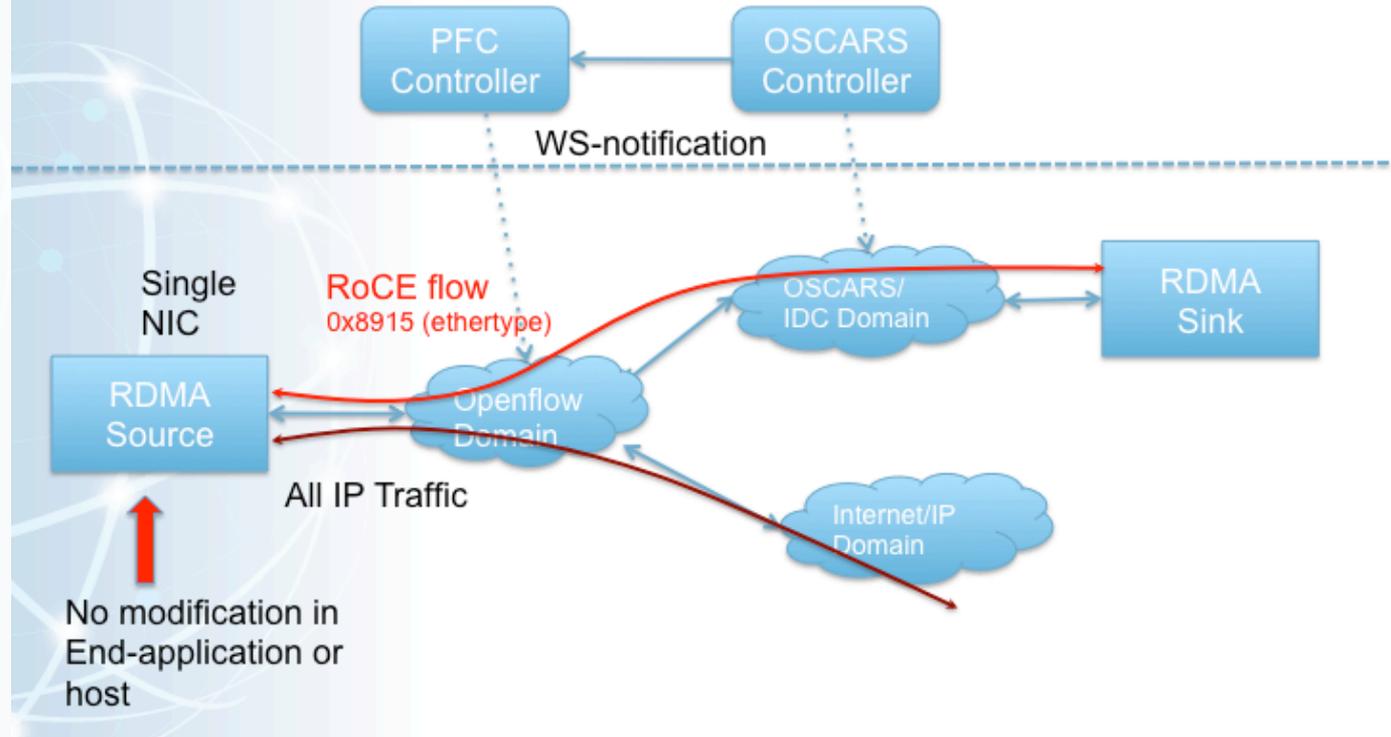


NEC

## Insights

- SDN not immune from end-to-end problem
- 'unmodified end host' an attractive architecture

## Demonstrating end-to-end RDMA flows



Lawrence Berkeley National Laboratory

U.S. Department of Energy | Office of Science

# Microsoft ONS 2014: Implemented our first idea in production



### SDN: Building the right abstractions for Scale

Abstract by separating management, control, and data planes

Example ACLs

Management plane	Create a tenant
Control plane	Plumb these tenant ACLs to these switches
Data plane	Apply these ACLs to these flows

- Data plane needs to apply per-flow policy to millions of VMs
- How do we apply billions of flow policy actions to packets?

### Flow Tables are the right abstraction

- VMSwitch exposes a typed Match-Action-Table API to the controller
- One table per policy
- Key insight: Let controller tell the switch exactly what to do with which packets (e.g. encaps/decap), rather than trying to use existing abstractions (Tunnels, ...)

### RDMA – High Performance Transport for Storage

- Remote DMA primitives (e.g. Read address, Write address) implemented on-NIC
  - Zero Copy (NIC handles all transfers via DMA)
  - Zero CPU Utilization at 40Gbps (NIC handles all packetization)
  - <2µs E2E latency
- RoCE enables Infiniband RDMA transport over IP/Ethernet network (all L3)
- Enabled at 40GbE for Windows Azure Storage, achieving massive COGS savings by eliminating many CPUs in the rack

All the logic is in the host:  
Software Defined Storage now scales with the Software Defined Network

### Just so we're clear... 40Gbps of I/O with 0% CPU

The screenshot shows the Windows Task Manager Performance tab. The CPU usage is highlighted in red and shows 0% utilization. The system is running on an Intel(R) Xeon(R) CPU E3-2690 0 @ 2.90GHz. Other system metrics like memory, ethernet, and disk usage are also visible.

# Journey with Programmability

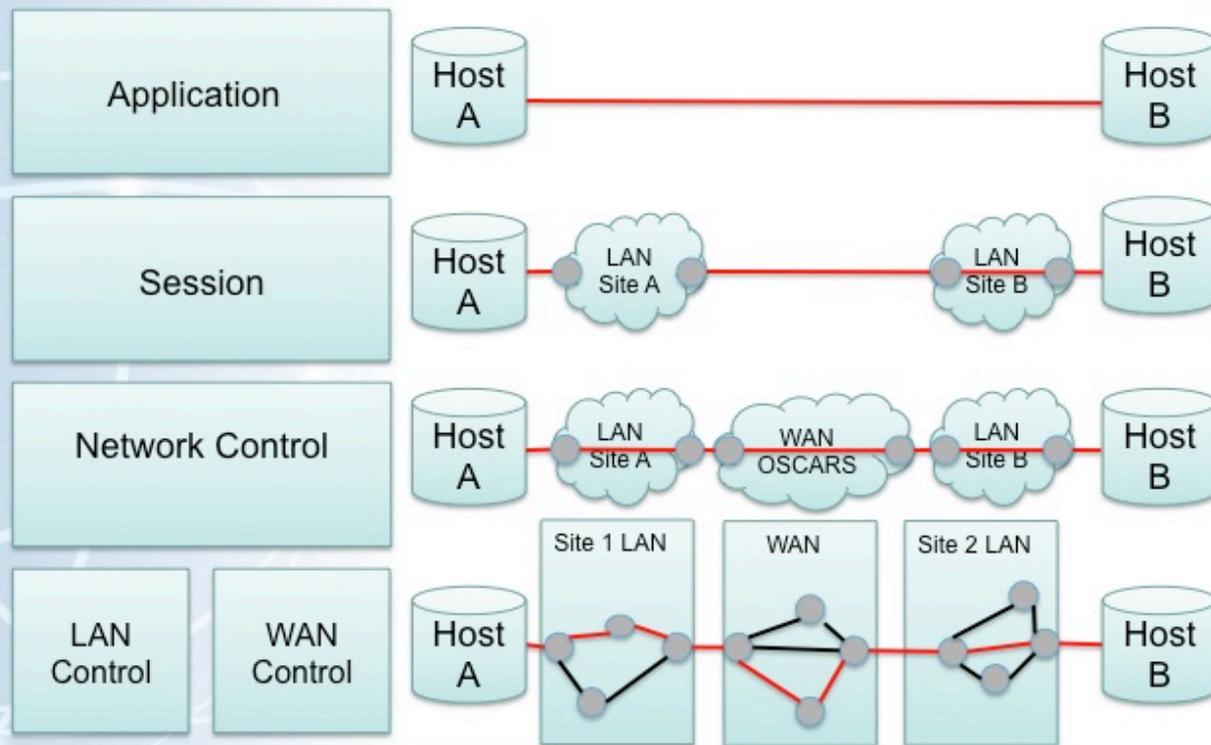
*Inaugural Open Network Summit, 2011 (Stanford)  
and SC 2011 (Seattle)*



## Insights

- **Zero configuration circuit**
- end-to-end communication at each layer important

## Brokering LAN and WAN Resources *a multi-layer view*



(even at layer 8)

7/17/12

Lawrence Berkeley National Laboratory

U.S. Department of Energy | Office of Science

# Journey with Programmability

SRS, Ciena, SuperComputing 2012, Salt Lake City



## Insights

- **Network virtualization is SDN 'killer app'**

- 'virtual switch' abstraction in the WAN holds promise

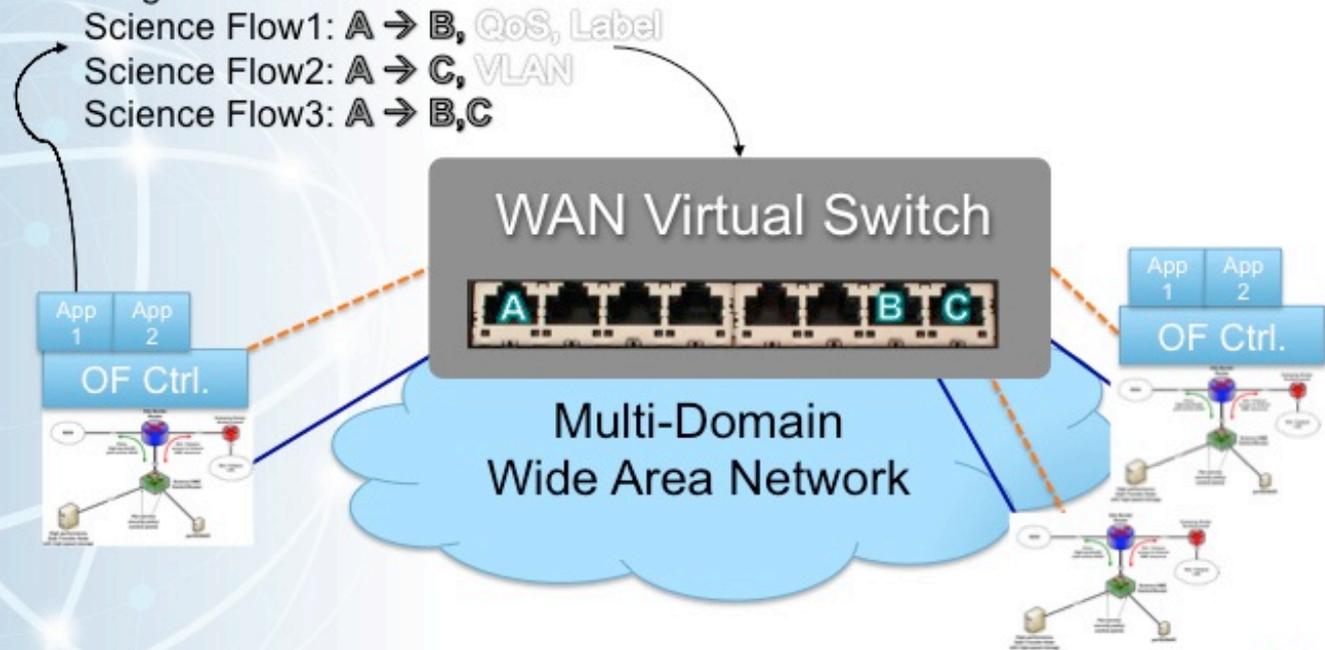
"Programmable" by end-sites

Program flows:

Science Flow1: A → B, QoS, Label

Science Flow2: A → C, VLAN

Science Flow3: A → B,C



# Journey with Programmability

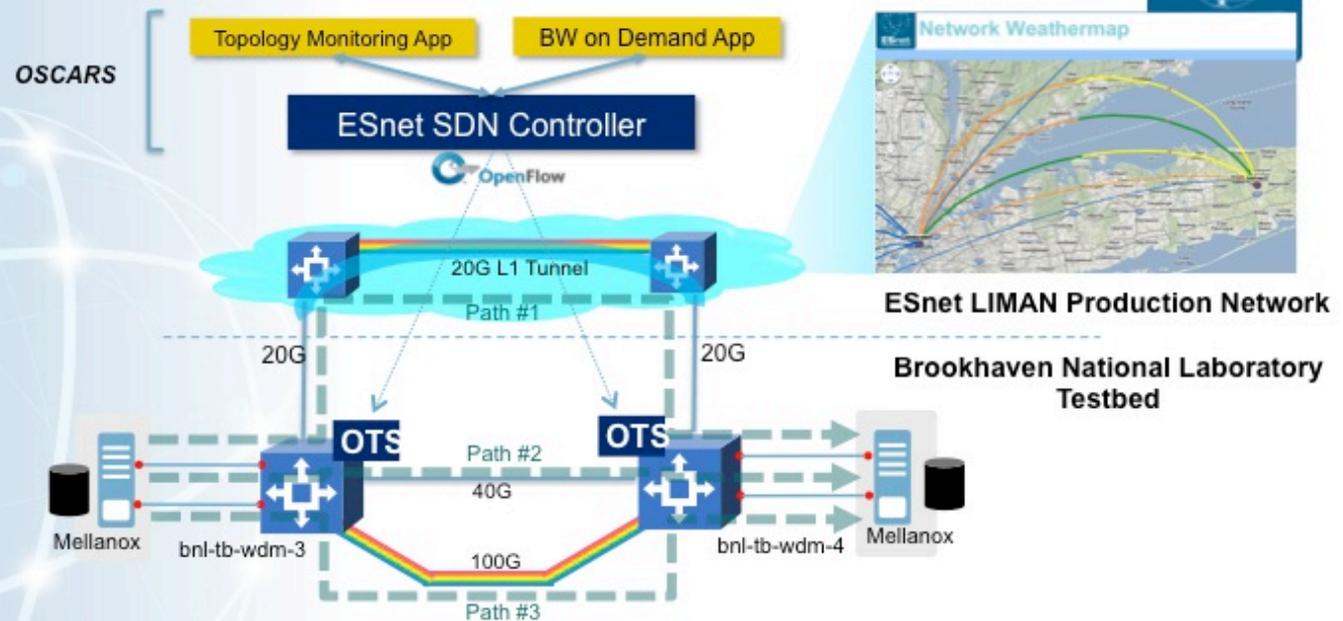
World's first Transport SDN Demo, Infinera/ESnet/Brookhaven  
December 2012



## Insights

- optical-layer automation essential for future topologies, architectures

## ESnet Transport SDN Demo



SDN Controller communicating with OTS via OpenFlow extensions

Bandwidth on Demand application for Big Data RDMA transport

3 physical transport path options (with varying latencies)

Implicit & explicit provisioning of 10GbE/40GbE services demonstrated



Lawrence Berkeley National Laboratory

U.S. Department of Energy | Office of Science

# Industry impact – led to the formation of Optical Transport Working Group (OTWG) in ONF



- Feasibility of using OpenFlow for controlling optical devices justified formalization of the OTWG in ONF
- Invited to be a ONF Research Associate
  - ONLY DOE/Lab person represented in the ONF (with 200 companies)
- Multiple companies working on implementing the changes – big demonstration in Germany in October timeframe
  - ESnet not participating in demonstration due to resource constraints

# Journey with Programmability

## BGP over SDN infrastructure, ONS 2013



### Front-Line Assembly

DEMO

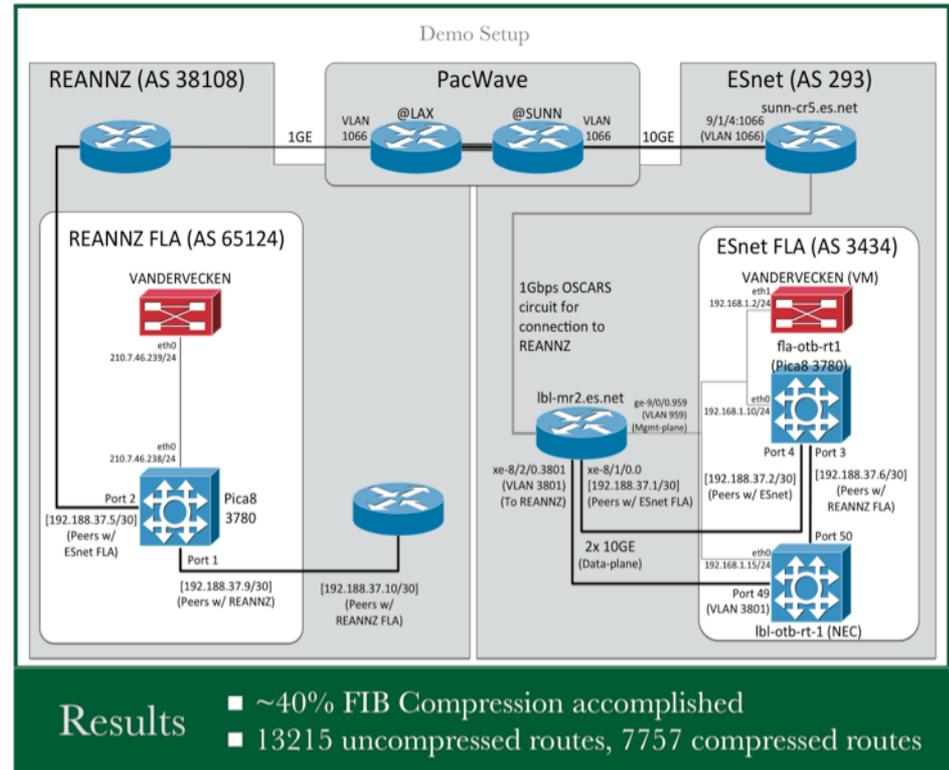
First international BGP peering using SDN in production between two national-scale network providers

- Innovative FIB compression enables using commodity OpenFlow switches for peering
- Leverages community open-source packages. RouteFlow and Quagga

### Insights

- SDN networks can interface with existing Internet
- New techniques need to be developed to scale controller based networking

Demonstration Team:  
 Google Network Research – Josh Bailey, Scott Whyte  
 REANNZ – Dylan Hall, Sam Russell, James Wix, Steve Cotter  
 ESnet – Inder Monga, Chin Guok, Eric Pouyoul, Brian Tierney  
 Acknowledgements - Joe Stringer



**Results**

- ~40% FIB Compression accomplished
- 13215 uncompressed routes, 7757 compressed routes

# Software-Defined Exchanges



## **SDX: A Software Defined Internet Exchange**

Nick Feamster<sup>†</sup>, Jennifer Rexford<sup>\*</sup>, Scott Shenker<sup>‡</sup>, Dave Levin<sup>◊</sup>, Russ Clark<sup>†</sup>, Josh Bailey<sup>\*</sup>

<sup>\*</sup> Google, <sup>†</sup> Georgia Tech, <sup>◊</sup> University of Maryland, <sup>\*</sup> Princeton University, <sup>‡</sup> UC Berkeley

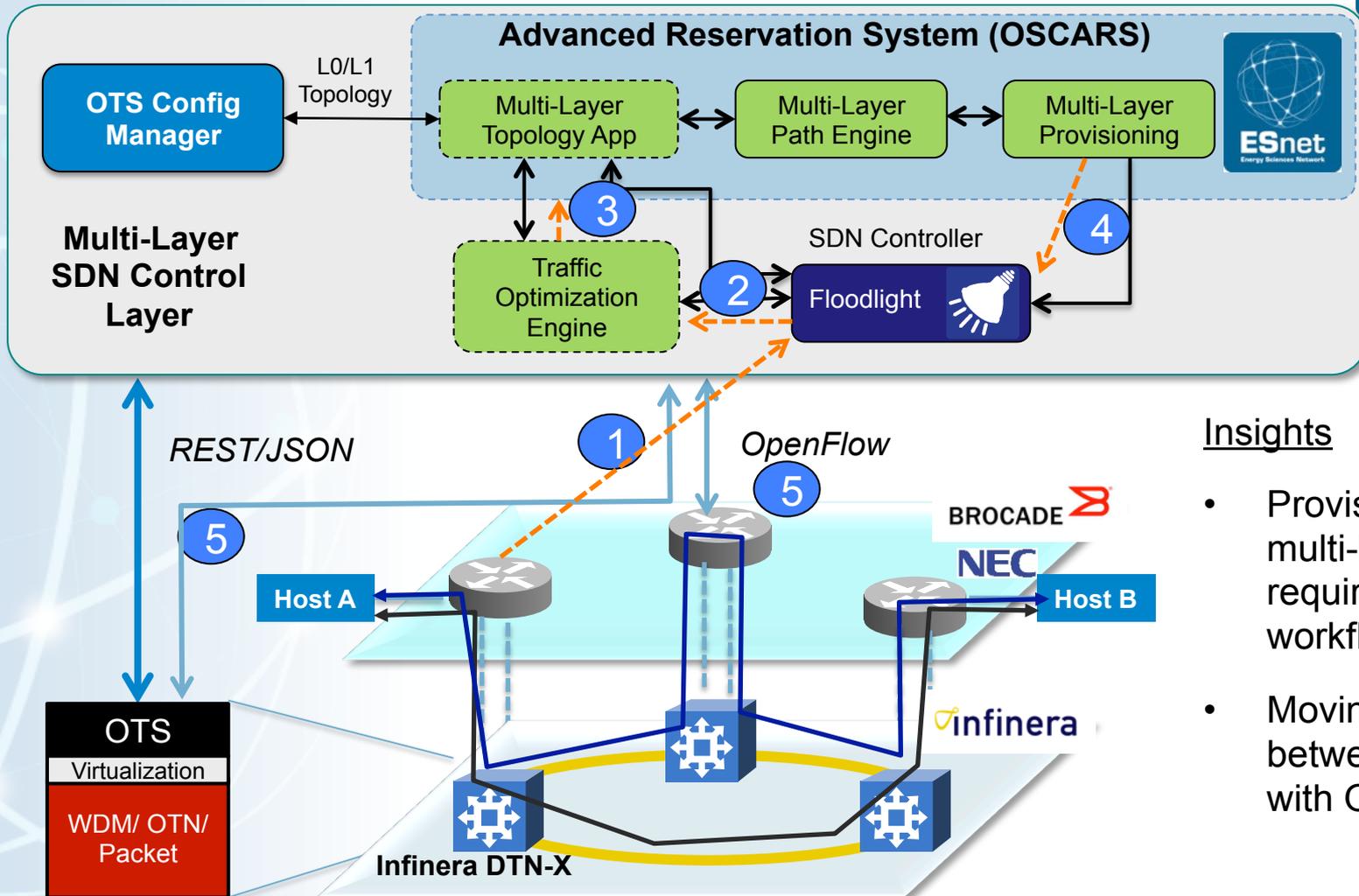
A lot of hype on this exchange concept

Allow multiple networks to build an exchange and apply application/  
network specific policy on traffic being exchanged

Definition being adapted by other folks to suit their purposes.

# Journey with Programmability

## Multi-layer SDN, Layer123 SDN, Oct 2013



### Insights

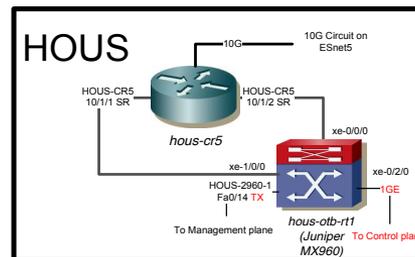
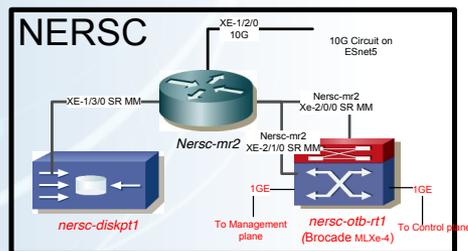
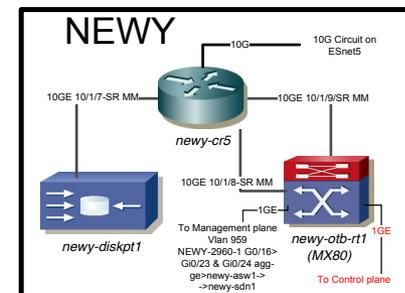
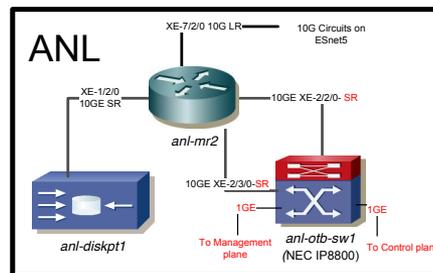
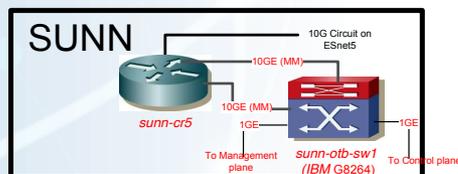
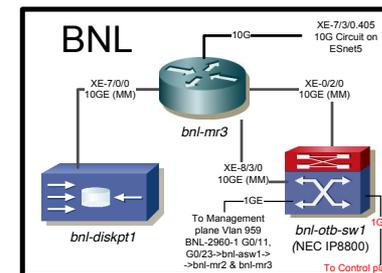
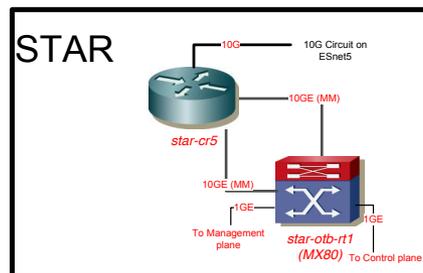
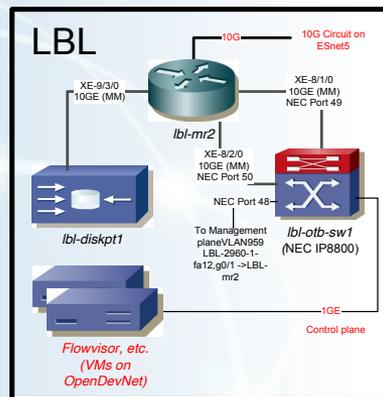
- Provisioning multi-layer could require a workflow
- Moving traffic between tunnels with OF is trivial

# SDN Testbed Final Plan

- not all the components are installed



## ESnet OpenFlow Testbed



Updated February 26, 2014



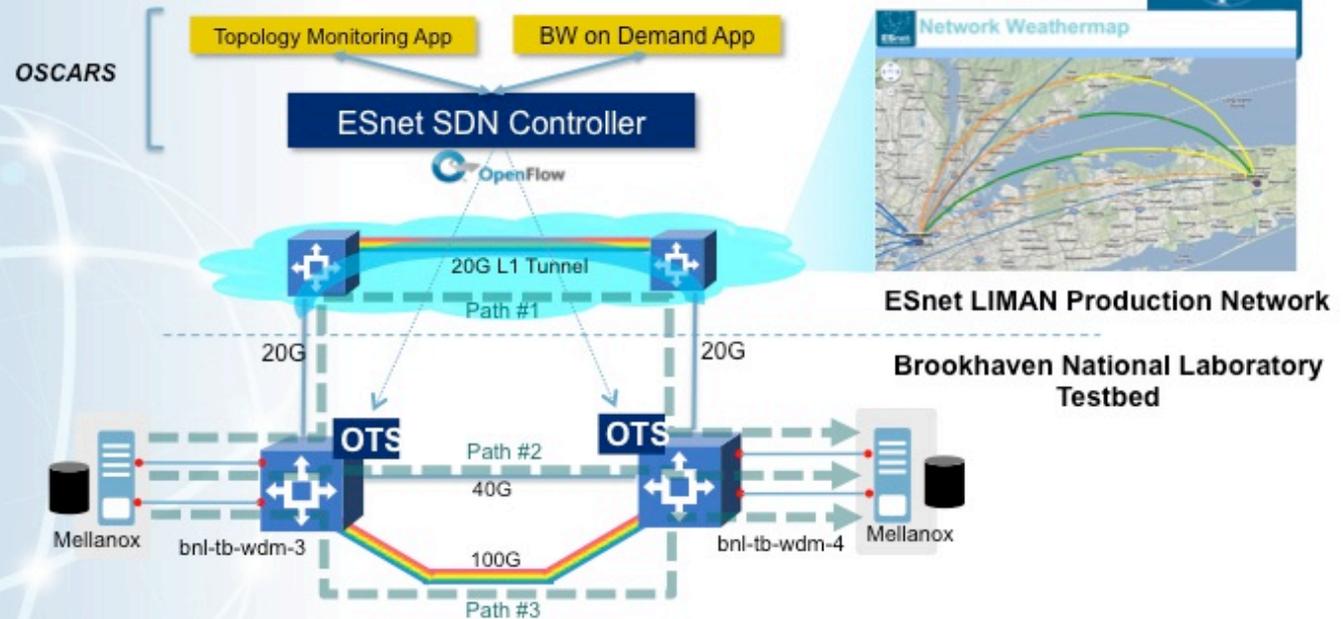
# OSCARS and SDN

# Journey with Programmability

World's first Transport SDN Demo, Infinera/ESnet/Brookhaven



## ESnet Transport SDN Demo



SDN Controller communicating with OTS via OpenFlow extensions

Bandwidth on Demand application for Big Data RDMA transport

3 physical transport path options (with varying latencies)

Implicit & explicit provisioning of 10GbE/40GbE services demonstrated



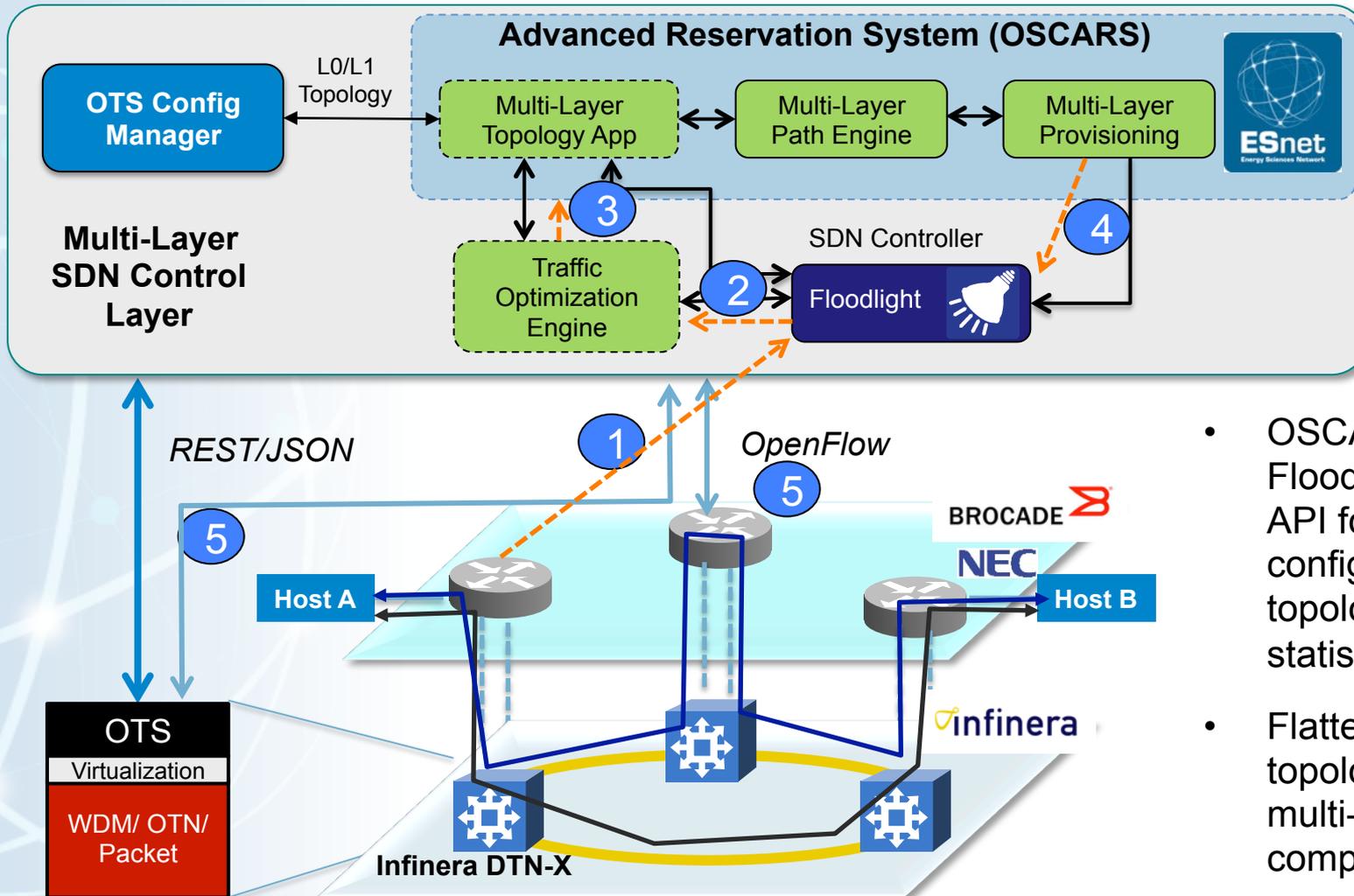
Lawrence Berkeley National Laboratory

U.S. Department of Energy | Office of Science

- OSCARS integrated OpenFlow as a PSS driver
- Vendor extensions to be compatible with OTS

# Journey with Programmability

## Multi-layer SDN, Layer123 SDN, Oct 2013



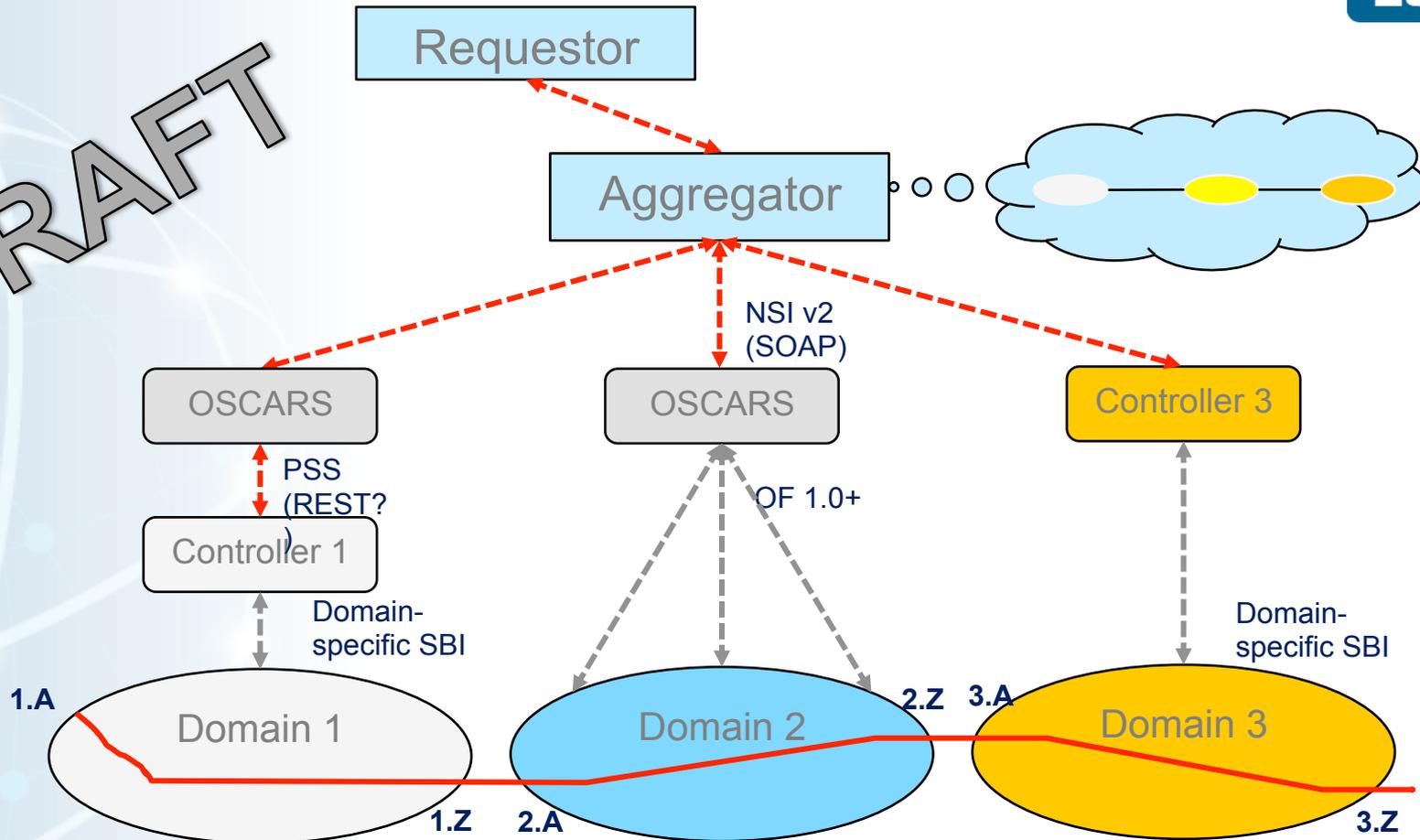
- OSCARS used Floodlight REST API for configuration, topology and statistics
- Flatten the topology for multi-layer path computation

# OGF/ESNet - Service Request NBI

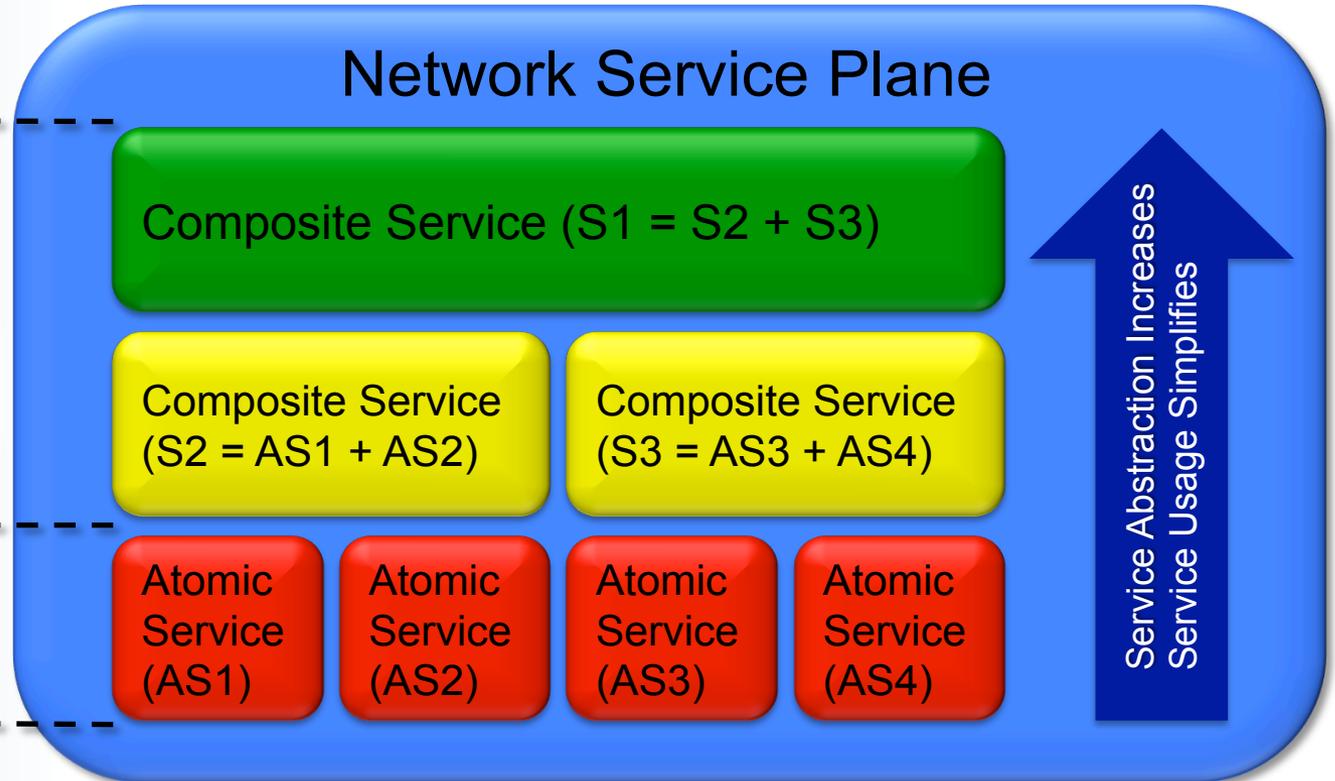
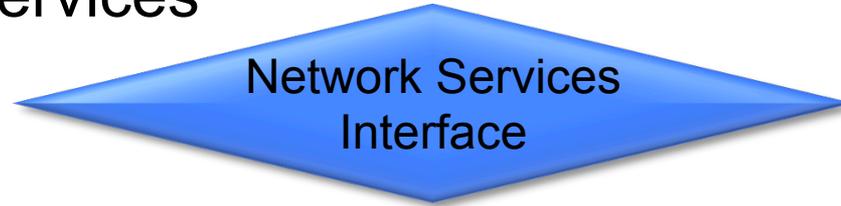
Allow simple service request across domain



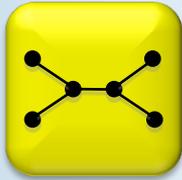
DRAFT



# Building Network Capabilities using Atomic and Composite Network Services



# Examples of Atomic Services



**Topology Service** to determine resources and orientation



**Routing Service** to enable IP connectivity



**Resource Computation Service** to determine possible resources based on multi-dimensional constraints



**Store and Forward Service** to enable caching capability in the network



**Connection Service** to specify data plane connectivity



**Measurement Service** to enable collection of usage data and performance stats



**Protection Service** to enable resiliency through redundancy



**Monitoring Service** to ensure proper support using SOPs for production service



**Restoration Service** to facilitate recovery

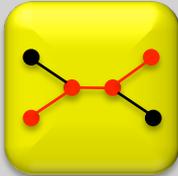


**Firewall Service** to prevent unwanted access of network resources

# Examples of Composite Network Services



## LHC: Resilient High Bandwidth Guaranteed Connection



Connection



Protection



Measurement



Monitoring

## L3VPN: Privately Routed Network (w/ Commodity)



Connection



Restoration



Routing



Firewall



Monitoring

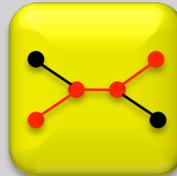
## Protocol Testing: Constrained Path Connection



Topology



Resource Computation



Connection



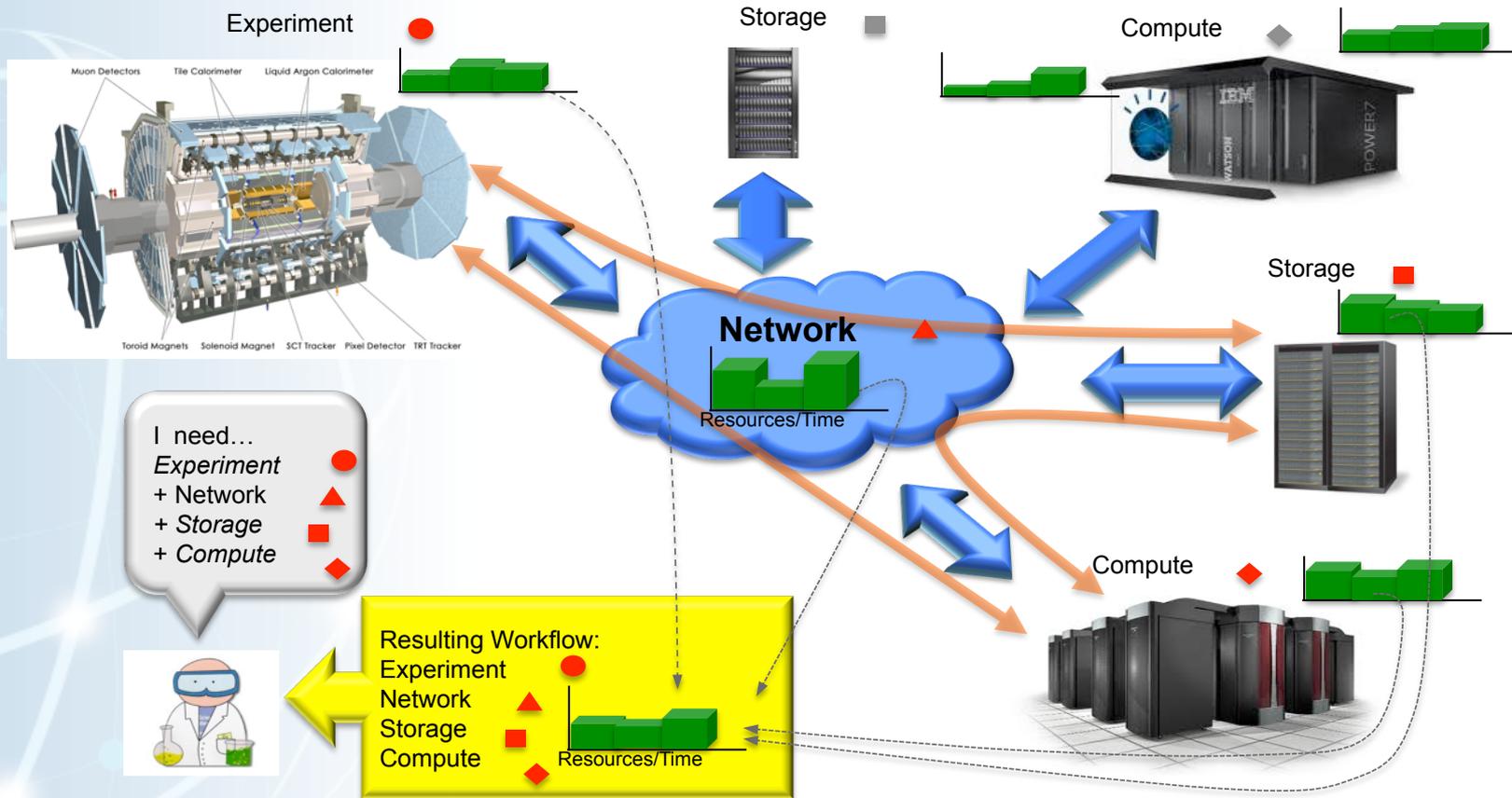
Measurement



# Conclusion: Application Workflow Integration is Critical!



A key focus is on technology development which allow networks to participate in application workflows



**The Network needs to be available to application workflows as a first class resource in this ecosystem**



# DOE/NSF Workshop 'SDN Operationalization'

# Three Goals

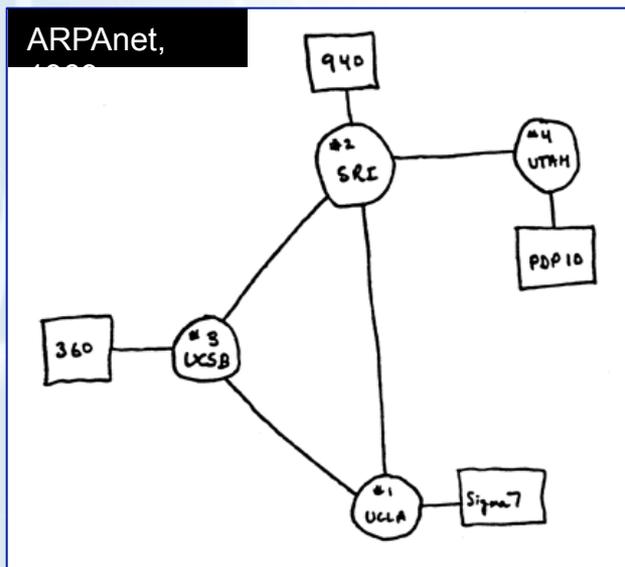


1. Bridge the 'operational' gap
  - architecture, tools and policies
2. Deploy and operate **securely** multi-layer, multi-domain SDN networks
  - Interwork with the current set of Internet technologies
3. Identify research, development and technologies needed to support new, innovative users and applications

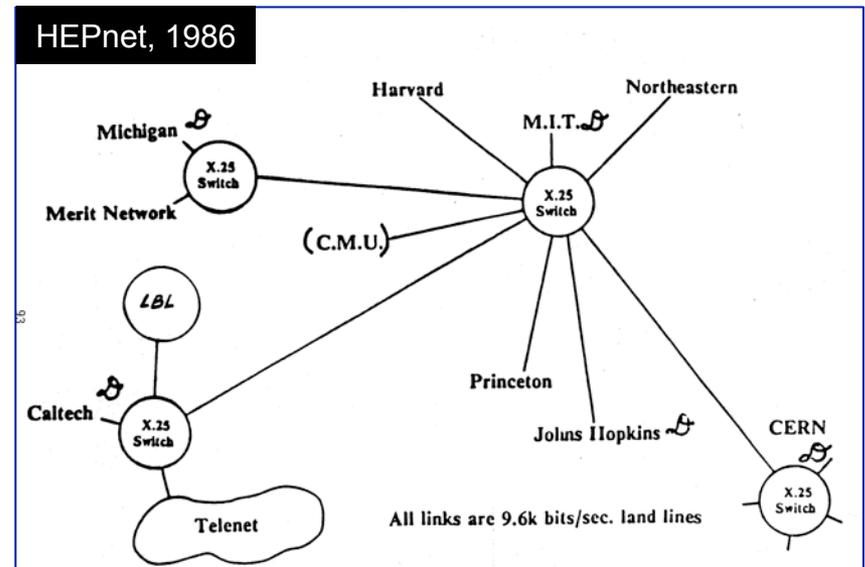
# It is about *running* networks (not just about network *research*)



Develop, deploy and (inter)operate  
a prototype multi-domain SDN network



[http://www.computerhistory.org/internet\\_history/full\\_size\\_images/1969\\_4-node\\_map.gif](http://www.computerhistory.org/internet_history/full_size_images/1969_4-node_map.gif)



Future of intersite networking, LBL, 1986

# Workshop to focus on the gaps and explore building a prototype multi-domain SDN network



1. Bridge the 'operational' gap
  - architecture, tools and policies
2. Deploy and operate **securely** multi-layer, multi-domain SDN networks
  - Interwork with the current set of Internet technologies
3. Identify research, development and technologies needed to support new, innovative users and applications

# Observations and gaps (only a subset)



## **Time is right for prototyping operational, multi-domain SDNs**

- Connect up testbeds and operational networks
- Build the SBone, the ARPAnet, ....with new technologies

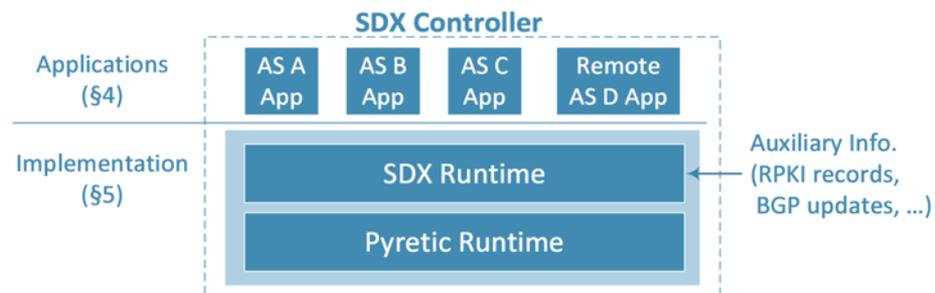
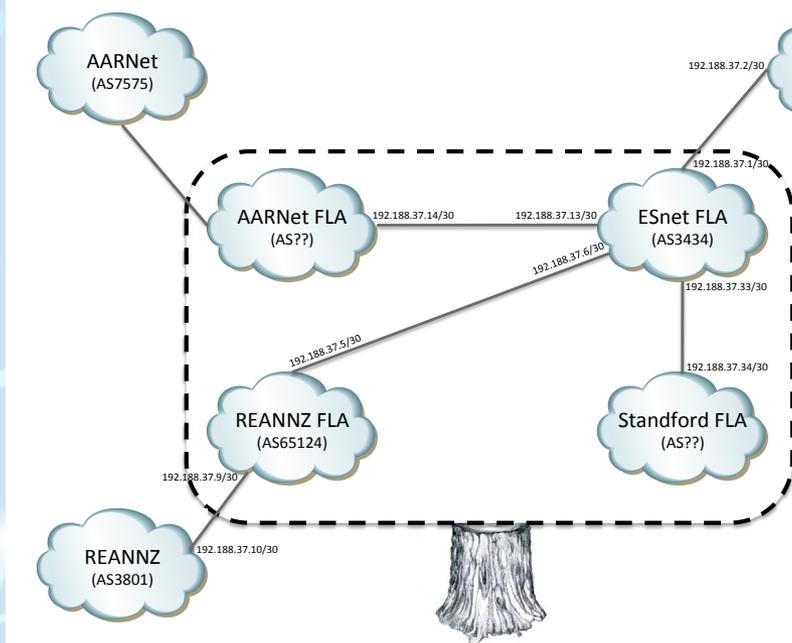
## **Multi-domain SDN deployment is key**

# Observations and gaps contd.

## Build multiple, community supported Software-Defined eXchanges (SDX)

- Tackling exchanging IP data is table-stakes for a larger deployment

TreeHouse Setup Overview [7/24/2013]



Collaboration with Josh Bailey, Google

Nick Feamster, Georial Tech